Sur la notion de moyenne d'une variable quantitative: clarifications théoriques et pratiques

Chénangnon F Tovissodé, Romain Glele-Kakaï

Sur la notion de moyenne d'une variable quantitative: clarifications théoriques et pratiques

Chénangnon F. Tovissodé [1] and Romain Glèlè Kakaï [1]

1 Laboratoire de Biomathématiques et d'Estimations Forestières, Université d'Abomey-Calavi, Benin

Email: romain.glelekakai@fsa.uac.bj

Résumé

Afin de guider les étudiants et les professionnels dans le choix du type de moyenne approprié pour une variable aléatoire, une définition opérationnelle de la valeur moyenne a été proposée dans Kpenavoun Chogou (2020). Cette note clarifie les sous-entendus de la définition opérationnelle proposée en partant de la définition mathématique formelle de la moyenne. La détermination de l'erreur type et d'un intervalle de confiance pour la moyenne, ainsi que les propriétés de la moyenne en échantillon fini sont décrites en complément.

Mots clés: moyenne, définition mathématique, type de moyenne, biais, erreur type

Abstract

In order to guide students and professionals in choosing the appropriate type of mean for a random variable, an operational definition of the mean has been proposed in Kpenavoun Chogou (2020). This working paper clarifies the implications of the operational definition proposed by starting from the formal mathematical definition of the mean. The computation of the standard error and of a confidence interval for the mean, as well as the properties of the finite sample mean are also described.

Keywords: mean, mathematical definition, type of mean, bias, standard error

Introduction

La moyenne est la statistique la plus utilisée pour résumer une variable ou une série statistique quantitative par un nombre, représentant la tendance centrale. Le calcul et l'utilisation de la moyenne sont donc très communs. Le concept de moyenne d'une variable remonte à l'antiquité chez les Egyptiens (plusieurs siècles avant J-C) et a été ensuite popularisée par les Grecs (Caveing, 1998). Il existe une grande multitude de moyennes dans la littérature (Bullen, 2013). Kpenavoun Chogou (2020) a récemment souligné la difficulté rencontrée par des professionnels de divers domaines ainsi que les étudiants dans le choix du type de moyenne à considérer dans différentes situations et a proposé une définition opérationnelle en vue de faciliter le calcul et l'identification du type de moyenne.

Cette note en complément de Kpenavoun Chogou (2020) apporte plus de clarification sur la notion de moyenne, l'importance de la moyenne arithmétique dans un contexte de modélisation et l'interprétation des valeurs moyennes. Elle décrit également la détermination du biais et de l'erreur type, ainsi que la construction d'un intervalle de confiance asymptotique autour d'une valeur moyenne calculée à partir d'un échantillon aléatoire. Des exemples sont donnés pour guider les professionnels qui produisent ou utilisent quotidiennement des valeurs moyennes.

1 Clarifications sur la notion de la moyenne

1.1 Définition naïve

Définition 1 Pour une variable aléatoire quantitative réelle, une valeur moyenne est un nombre réel pouvant s'écrire comme une combinaison des valeurs possibles de la variable et qui indique une tendance centrale de cette variable.

Dans une population, une moyenne d'un caractère est intégralement définie par la donnée de sa distribution. En revanche, sur la base d'une série statistique donnée (*i.e.* un échantillon), une moyenne est une statistique, *i.e.* une fonction des éléments de la série. L'interprétation d'une valeur moyenne calculée dépend alors de l'usage envisagé: la caractérisation de la série statistique, ou la caractérisation de la population-mère dont provient la série (inférence).

Il est évident que la définition naïve est très vague du point de vue quantitatif car elle n'est

explicite ni sur la signification ni sur la méthode de calcul de la moyenne. Sans aucune précision et sans élément contextuel explicite, le terme *moyenne* fait référence à la moyenne arithmétique. Cette dernière est la valeur moyenne la plus utilisée, probablement parce qu'elle est dans une population la valeur attendue (*i.e.* l'espérance mathématique) et en échantillon fini, un estimateur *non biaisé et consistent* de la valeur attendue. Il existe néanmoins plusieurs types de moyennes. Entre autres indicateurs basiques de tendance centrale, on peut citer les moyennes mobiles (Goldfarb, 2011) et les moyennes arithmétiques trimées qui écartent les valeurs extrêmes et donnent des tendances centrales robustes aux valeurs aberrantes ou extrêmes (Prescott, 1978). Ces moyennes sont basiques dans le sens qu'elles n'impliquent aucune transformation de la variable d'intérêt. Notons qu'une mesure basique donnée peut être indéterminée pour certaines distributions, *e.g.* l'espérance mathématique d'une variable de loi de Cauchy. Certaines quantiles, notamment la médiane sont parfois utilisés comme valeurs centrales, mais ne sont pas des moyennes. Cependant, pour certaines distributions (*e.g.* double exponentielle), la médiane d'un échantillon est un meilleur estimateur de l'espérance mathématique que la moyenne arithmétique. Au sens étendu, une moyenne peut être plus formellement définie comme dans la sous-section suivante.

1.2 De la nécessité de généralisation de la notion de moyenne: définition mathématique

Définition 2 Étant donnée une variable aléatoire réelle X, un opérateur $C[\cdot]$ qui donne une mesure basique de tendance centrale (*e.g.* l'espérance mathématique, une espérance trimée), une transformation monotone $\psi(\cdot)$ et sa réciproque $\psi^{-1}(\cdot)$; une moyenne de la variable X est donnée par

$$m = \psi^{-1}((C[\psi(X)])\,. \tag{1}$$

En générale, la mesure basique $C[\cdot]$ utilisée est l'*espérance mathématique* dont l'opérateur est ici noté $E[\cdot]$. On parle alors de moyenne *quasi-arithmétique*, ou de *moyenne régulière* lorsque $\psi(x) \neq x$ (de Carvalho, 2016).

Partant de (1), Kolmogorov (1930) a introduit la définition de la moyenne la plus utilisée et unanimement acceptée par les statisticiens tout au moins pour les moyennes quasi-arithmétiques.

Elle considère la moyenne comme étant simplement une fonction interne de régularisation numérique. Une moyenne m d'une variable aléatoire X est dans ce cas (Kolmogorov, 1930):

$$m = \psi^{-1}(E[\psi(X)]). \tag{2}$$

Dans un échantillon fini $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ (aléatoire simple), en posant $E[\psi(X)] = T_o = \frac{1}{n}\sum_{i=1}^{n}\psi(x_i)$, la moyenne s'identifie à

$$\widehat{m} = \psi^{-1}\left(T_o\right). \tag{3}$$

Dans un échantillon aléatoire en générale, la statistique T_o prend la forme $T_o = \sum_{i=1}^{n} p_i \psi(x_i)$ avec $\sum_{i=1}^{n} p_i = 1$, p_i étant le poids relatif de la valeur x_i. Les moyennes au sens étendu les plus populaires sont la moyenne quadratique ($\psi(t) = t^2$ pour $t > 0$ et $\psi^{-1}(x) = \sqrt{x}$), la moyenne géométrique ($\psi(t) = \log(t)$ et $\psi^{-1}(x) = \exp(x)$ pour $x > 0$) et la moyenne harmonique ($\psi(t) = \psi^{-1}(t) = t^{-1}$ pour $t \neq 0$). D'autres types de moyenne relativement peu utilisés peuvent être aussi considérés: la moyenne-puissance ($\psi(t) = t^{\alpha}$ et $\psi^{-1}(x) = x^{\frac{1}{\alpha}}$ pour $x > 0$) et la moyenne exponentielle ($\psi(t) = \exp(\alpha t)$ et $\psi^{-1}(x) = \frac{1}{\alpha}\log(x)$ pour $x > 0$). On déduit aisément que pour la moyenne arithmétique, $\psi(t) = \psi^{-1}(t) = t$. En dehors de la moyenne arithmétique, tous les autres types de moyenne définis ci-dessus sont dit *quasi-arithmétique*.

1.3 Lien entre la signification et le calcul d'une moyenne

De (2), on note que la seule mention de la moyenne ne suffit pas à définir une tendance centrale au sens étendu. Le calcul d'une valeur moyenne nécessite la précision de la transformation $\psi(\cdot)$ en (1). Ainsi, afin d'éviter toute ambiguïté, la mention de la moyenne devrait au sens étendu être accompagnée d'une indication de la fonction $\psi(\cdot)$. Une telle indication peut être explicite: préciser un type de moyenne commune (arithmétique, géométrique, quadratique, harmonique). Elle peut également être implicitement donnée à travers la précision de la signification exacte de la moyenne désirée.

En fait, dans n'importe quel contexte particulier, une précision de la signification d'une valeur moyenne serait suffisante pour une définition sans équivoque. Ainsi, dans l'étude des peuplements forestiers, la mesure centrale de diamètre généralement considérée est la moyenne quadratique des diamètres. Une quantité importante dans ce domaine est en fait la surface terrière, *i.e.*

l'aire de la section transversale d'un arbre à 1,3 m du sol (proportionnelle au carré du diamètre de l'arbre) qui exprimée en m^2/ha permet de visualiser la surface effectivement occupée par les arbres sur un site (Rondeux, 1978). Une approche simple mais précise est de définir (et d'interpréter) cette moyenne quadratique comme le *diamètre de l'arbre de surface terrière moyenne* (*i.e.* attendue).

Une situation similaire apparait en considérant une entreprise transportant des boules d'une même densité mais de diamètre variable. Le volume occupé par une boule est une quantité importante à considérer pour les prévisions de chargement. Dans un tel contexte, le diamètre de la boule de volume moyen est une mesure de tendance centrale appropriée. La moyenne des diamètres est dans ce cas définie comme une moyenne quasi arithmétique avec une transformation $\psi(t) = t^3$ et sa réciproque $\psi^{-1}(x) = x^{1/3}$.

Dans les sciences actuaires, les taux de change sont d'une importance capitale et la moyenne généralement considérée pour cette quantité est la moyenne harmonique. En fait, dans ce contexte, les quantités primaires d'intérêt sont les montants transférés et les montants reçus. La moyenne harmonique des taux de change est dans ce contexte le taux de change permettant de convertir l'espérance mathématique du montant transféré en l'espérance mathématique du montant reçu (*i.e.* convertir le montant moyen transféré en montant moyen reçu). Une telle indication permettrait à tout professionnel/étudiant de comprendre sans ambiguïté le sens de "taux de change moyen", ce qui permettrait d'éviter les erreurs commises. Dans Kpenavoun Chogou (2020), deux exemples de séries de taux de change de versements réalisés par la Coopération Suisse à deux différents laboratoires sont considérés. Dans le premier cas (Tableau 3), on dispose de taux de change (1 Franc Suisse en FCFA) et des montants reçus par le laboratoire concerné (en FCFA). La moyenne des taux d'intérêt est dans ce cas, identifiée à une moyenne harmonique des taux individuels. Dans le second cas (Tableau 4), on dispose des taux de change et des montants transférés par la Coopération Suisse (en Francs Suisse). La moyenne des taux d'intérêt est ici identifiée à une moyenne arithmétique des taux pondérés par les montants transférés. L'auteur note à juste titre que des problèmes d'interprétation se posent lorsque le type de moyenne approprié n'est pas considéré (mauvaise spécification de la fonction ψ dans (2)). Toutefois, une mauvaise spécification de la fonction ψ pose plus

un problème d'interprétation de la moyenne qu'une erreur de calcul, la valeur obtenue d'une moyenne arithmétique pouvant être considéréeà la place d'autres types de moyenne dans certaines situations.

Par ailleurs, partant de (2), une définition litérale de la notion de moyenne pourra difficillement prendre en compte toutes les spécificités liées à la notion de moyenne, d'où la difficulté d'une généralisation de la notion de moyenne au sens litérale par les livres de statistiques et considérée à tord comme une insuffisance par Kpenavoun Chogou (2020) (Page 3). Comme illustration, la définition opérationnelle d'une moyenne proposée par Kpenavoun Chogou (2020) est certes nécessaire comme préalable au calcul d'une moyenne mais n'est pas suffisante. En d'autres termes, la proposition inverse n'est pas vraie, donc cette définition opérationnelle est tout aussi naïve. En effet, elle indique que, pour une variable quantitative réelle, la valeur moyenne "*est une mesure de tendance centrale ayant la propriété de conserver la caractéristique de l'ensemble des observations quand on remplace chacune de ces observations par cette valeur unique.*" D'abord, d'un point de vue fondamental, la moyenne n'est pas une mesure mais une fonction. De plus, nombre de fonctions n'ayant aucun rapport avec la moyenne ont pourtant la même définition opérationnelle. Un exemple concret est la fonction g de Ricci (1915) définie comme suit:

$$\forall(x_1, x_2, ..., x_n), \quad g(x_1, x_2, ..., x_n) = x_n + (x_n - x_1) + (x_n - x_2) + ... + (x_n - x_{n-1}).$$

Cette fonction vaut x_n lorsque $x_1 = x_2 = ... = x_n$, pourtant, elle n'est pas une fonction de moyenne. De manière plus spécifique, en considérant par exemple x_n comme la médiane de la série, la fonction retournera toujours la médiane de cette série qui est différente de la moyenne.

1.4 Pondération et pseudo moyennes pondérées

La pondération d'une moyenne d'une série statistique vise la prise en compte du plan d'échantillonnage utilisé pour extraire la série de la population-mère. En d'autres termes, le poids associé à chaque valeur d'une série statistique indique la proportion d'individus de la population qu'elle représente dans l'échantillon. Dans la définition (1), la pondération est prise en compte par l'opérateur $C[\cdot]$ qui donne une mesure basique de tendance centrale.

Dans une série statistique issue d'un échantillonnage aléatoire et simple de taille n (tous les individus de la population ont la même probabilité d'être sélectionné), les valeurs de la série sont

auto-pondérées, *i.e.* chaque valeur dans chaque échantillon représente $\frac{100}{n}\%$ de la population. Considérons une population constituée de K strates $Str_1, Str_2, \cdots, Str_K$ (une strate étant une sous-population plus homogène par rapport à la population globale), avec p_k la proportion d'individus dans la strate Str_k. Un sous-échantillon de taille n_k est extraite de chaque strate Str_k pour former un échantillon global de taille $n = n_1 + n_k + \cdots + n_K$. Dans ce cas, chaque valeur provenant de la sous-population Str_k représente dans l'échantillon $\frac{100 \times p_k}{n_k}\%$ de la population totale. Ainsi, si dans le cadre d'un échantillonnage stratifié, on extrait de chaque strate Str_k un échantillon de taille $n_k = p_k \times n$, alors les éléments de la série statistique de taille n obtenue sont auto-pondérés. Dans la cas contraire, tout calcul de moyenne basique doit prendre en compte les poids $\frac{p_k}{n_k}$ des différents éléments de la série.

Les poids des éléments d'une série ne sont pas nécessairement des nombres entiers. Considérons par exemple l'estimation du rendement moyen d'une spéculation à l'échelle nationale à partir d'une série de rendements observés sur des parcelles. L'effort d'échantillonnage est la somme S des superficies s_k des parcelles considérées et chaque parcelle représente $\frac{100 \times s_k}{S}\%$ de la campagne. Le poids associé au rendement de chaque parcelle est la superficie de la parcelle divisée par la superficie totale ($\frac{s_k}{S}$).

Il est important de noter que la moyenne est une caractéristique d'une unique variable aléatoire, *i.e.* la moyenne est une statistique univariée. La détermination d'une tendance centrale d'une variable ne requiert donc pas la connaissance de la relation mathématique entre cette variable et tout autre quantité (aléatoire ou non) en dehors des probabilités (ou densité de probabilité) associées aux différentes valeurs que peut prendre la variable.

Il convient ici de distinguer des pseudo moyennes pondérées définies dans des domaines de recherche et ou d'application spécifiques sur la base de deux ou plusieurs variables aléatoires. Pour exemple, considérons la "hauteur moyenne de Lorey" d'un peuplement forestier, définie comme la moyenne arithmétique des hauteurs totales des arbres, chacun pondéré par sa surface terrière (Rondeux, 1978). La hauteur moyenne de Lorey n'est clairement pas une simple mesure de tendance centrale de la hauteur des arbres. Si on définie le volume commercial (ou aérien) d'un arbre comme la moitié du produit de sa surface terrière par sa hauteur totale (formule de Huber (Rondeux, 1978)), la hauteur moyenne de Lorey représenterait alors la hauteur correspondante à

un arbre de volume commercial moyen (attendu) et de surface terrière moyenne (attendue). Un deuxième exemple concerne les indices des prix et des quantités de *Paasche* défini chacun pour un ensemble de biens comme une moyenne harmonique pondérée par les valeurs monétaires totales (quantité multipliée par prix) des différents biens considérés (Goldfarb, 2011). Clairement, la seule connaissance de la série des prix ne permet pas de calculer l'indice des prix de *Paasche*. Étant donné que le poids est ici une variable aléatoire au même titre que le prix et la quantité, les indices sont des pseudo moyennes harmoniques pondérées.

D'autres classifications des moyennes ont été proposées notamment les moyennes Lagrangiennes et les moyennes de Cauchy dont une description est faite dans Marichal (2006).

2 Observations sur l'utilisation de la moyenne arithmétique

2.1 La moyenne arithmétique a toujours un sens

L'espérance mathématique d'une variable aléatoire est la valeur attendue pour une réalisation unique. Comme indiqué dans § 1.1, la moyenne arithmétique est un estimateur non biaisé et consistent de la valeur attendue d'un caractère statistique (voir aussi § 3.2). Considérons par exemple les données présentées au tableau 7 dans Kpenavoun Chogou (2020). Il s'agit des coefficients de revalorisation quadri-annuel du salaire d'un fonctionnaire d'une entreprise entre 1985 et 2017. Si la valeur centrale appropriée pour décrire la variation relative du salaire (voir justification ici 4.1) sur la période considérée est la moyenne géométrique des coefficients individuels ($\bar{X}_g = 1{,}140$, soit une variation quadri-annuel de 14%), la moyenne arithmétique ($\bar{X} = 1{,}176$ soit une variation quadri-annuel de 17,6%) des coefficients est une meilleure prédiction pour la prochaine revalorisation du salaire du fonctionnaire en 2021.

La moyenne arithmétique n'est ainsi jamais dépourvue de sens. Elle peut être présentée, conjointement avec tout autre type de moyenne, comme statistique descriptive et interprétée comme valeur attendue.

2.2 Modèles statistiques et moyenne arithmétique

La modélisation statistique vise la prédiction d'une variable aléatoire. La prédiction d'une variable consiste à déterminer conditionnellement à des covariables, l'espérance mathématique

de cette variable. Après l'ajustement d'un modèle, la valeur attendue de la variable modélisée est obtenue comme moyenne marginale (aussi appelée "least square mean") du modèle. Ce principe de modélisation justifie l'utilisation généralisée de la moyenne arithmétique comme mesure de tendance centrale privilégiée lors d'une analyse statistique descriptive. En effet, l'utilisation de la moyenne arithmétique comme valeur centrale permet une comparaison entre la moyenne observée et la valeur attendue pour ainsi apprécier l'apport des covariables utilisées dans le modèle ajusté.

3 Inférence sur les moyennes quasi-arithmétiques

3.1 Biais et précision d'une moyenne

Lorsqu'une valeur moyenne $\widehat{m}$ est calculée sur un échantillon aléatoire pour servir de mesure de tendance centrale pour toute une population, la détermination de $\widehat{m}$ est une *estimation* (au sens fréquentiste) de la vraie moyenne m. On parle alors d'*inférence* statistique sur la moyenne. L'expression (3) de $\widehat{m}$ est un estimateur de m et la valeur $\widehat{m}$ est une *moyenne estimée*, *i.e.* une estimation ponctuelle de m.

En général, on se base sur certaines qualités des estimateurs pour juger de leur performance. Un estimateur est de bonne qualité lorsqu'il est consistant, *non biaisé* et de bonne précision. La consistance se réfère en pratique à la propriété d'un estimateur dont la valeur estimée du paramètre tend vers la valeur vraie lorsque la taille de l'échantillon tend vers l'infini (la loi faible des grand nombres). Les deux propriétés des moyennes qui nous intéressent ici sont le biais et la précision. Un estimateur $\hat{\theta}_n$ d'une caractéristique θ est *sans biais* ou *non biaisé* lorsque son espérance mathématique est égale θ: $\mathbb{E}(\hat{\theta}_n) = \theta$. Lorsque l'estimateur est biaisé, le biais $\mathbb{B}(\hat{\theta}_n, \theta)$ est tel que: $\mathbb{B}(\hat{\theta}_n, \theta)$=$\mathbb{E}(\hat{\theta}_n) - \theta$. Il est utile de noter ici que, parmi les différents types de moyenne décrits plus haut, seule la moyenne arithmétique est non biaisée. Toutes les moyennes quasi-arithmétiques sont biaisées et de ce fait, ne peuvent être calculées sans une correction tenant compte de leur biais. Par ailleurs, la précision d'un estimateur se réfère à l'erreur quadratique attendue (EQA), $\mathbb{E}[(\hat{\theta}_n - \theta)^2]$ prenant en compte la dispersion des observations autour de la moyenne, $V(\hat{\theta}_n)(= \sigma_o^2)$, et l'erreur systématique liée à son biais, $[\mathbb{E}(\hat{\theta}_n) - \theta]^2$:

$$EQA = \sigma_o^2 + [\mathbb{E}(\hat{\theta}_n) - \theta]^2.$$

Il est donc inutile de disposer d'une valeur estimée sans une mesure d'incertitude (Palm, 2002). Pour un estimateur $\widehat{m}$ de m, l'incertitude se compose donc du biais et de la précision.La sous-section suivante décrit comment obtenir des approximations du biais, de l'erreur type (mesure de précision) et un intervalle de confiance sur une estimation $\widehat{m}$ (3) de la valeur moyenne m (1) d'une variable aléatoire réelle X, à partir d'un échantillon de n observations $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$. Seules la moyenne arithmétique et les moyennes quasi-arithmétiques sont considérées. En d'autres termes, seules les moyennes utilisant l'espérance mathématique $E[\cdot]$ comme mesure basique de tendance centrale sont traitées. La détermination des expressions fournies est présentée en annexe.

3.2 Correction du biais et précision de moyennes (quasi) arithmétiques

Le Tableau 1 présente les formules des biais et des erreurs types des moyennes arithmétique et quasi-arithmétiques communes (quadratique, géométrique et harmonique) ainsi que les expressions pour corriger ces biais ($\widehat{m}_c$). En dehors de la moyenne arithmétique qui est un estimateur non-biaisé de l'espérance mathématique ($\widehat{m}_c = \widehat{m}$), les moyennes quasi-arithmétiques communément calculées sont biaisées et ne reflètent donc pas les moyennes correspondantes dans la population. En particulier, la moyenne quadratique d'un échantillon sous-estime toujours la moyenne quadratique de la population alors que la moyenne géométrique d'un échantillon sur-estime toujours la moyenne géométrique de la population. De même, la moyenne harmonique d'un échantillon d'une quantité positive (respectivement négative) sur-estime (respectivement sous-estime) la moyenne harmonique de la population.

Pour exemple, revenons sur le coefficient de revalorisation quadri-annuel du salaire d'un fonctionnaire: {1,012; 1,013; 1,006; 1,017; 1,023; 1,003; 1,006; 2,008; 1,501} (Tableau 7) dans Kpenavoun Chogou (2020). La moyenne géométrique est $\widehat{m} = 1{,}140$. La moyenne géométrique corrigée pour le biais est donnée par $\widehat{m}_c = \left(1 - 0.5\widehat{\sigma_o^2}\right)\widehat{m}$ et l'erreur type est $\widehat{\sigma}_{\widehat{m}} = \widehat{m}\widehat{\sigma}_o$ avec $\widehat{\sigma_o^2}$ la variance logarithmique estimée pour la population des coefficients: $\widehat{\sigma_o^2} = 0{,}007$. On obtient ainsi un coefficient moyen de $\widehat{m}_c = 1{,}137$ (soit une variation relative quadri-annuel du salaire de 13,7%) avec une erreur type de $\widehat{\sigma}_{\widehat{m}} = 0{,}095$.

Table 1. Estimateurs communs ($\widehat{m}$), biais à l'ordre 2 ($Biais(\widehat{m})$), estimateurs corrigés ($\widehat{m}_c$) et erreurs types à l'ordre 1 ($\sigma_{\widehat{m}}$) de moyennes quasi-arithmétiques communes

Moyenne	T_o	$\widehat{\sigma_o^2}$	$\widehat{m}$	$Biais(\widehat{m})$	$\widehat{m}_c$	$\widehat{\sigma}_{\widehat{m}}$
Arithmétique	$\frac{1}{n}\sum_{i=1}^n x_i$	$\frac{1}{n(n-1)}\sum_{i=1}^n (x_i - T_o)^2$	T_o	0	$\widehat{m}$	$\widehat{\sigma}_o$
Quadratique	$\frac{1}{n}\sum_{i=1}^n x_i^2$	$\frac{1}{n(n-1)}\sum_{i=1}^n (x_i^2 - T_o)^2$	$\sqrt{T_o}$	$-\frac{\widehat{\sigma_o^2}}{8\widehat{m}^3}$	$\widehat{m} + \frac{\widehat{\sigma_o^2}}{8\widehat{m}^3}$	$\frac{\widehat{\sigma}_o}{2\widehat{m}}$
Géométrique	$\frac{1}{n}\sum_{i=1}^n \log(x_i)$	$\frac{1}{n(n-1)}\sum_{i=1}^n (\log x_i - T_o)^2$	e^{T_o}	$\frac{1}{2}\widehat{m}\widehat{\sigma_o^2}$	$\widehat{m} - \frac{\widehat{m}}{2}\widehat{\sigma_o^2}$	$\widehat{m}\widehat{\sigma}_o$
Harmonique	$\frac{1}{n}\sum_{i=1}^n x_i^{-1}$	$\frac{1}{n(n-1)}\sum_{i=1}^n (x_i^{-1} - T_o)^2$	T_o^{-1}	$\widehat{m}^3\widehat{\sigma_o^2}$	$\widehat{m} - \widehat{m}^3\widehat{\sigma_o^2}$	$\widehat{m}^2\widehat{\sigma}_o$

Comme deuxième illustration, reconsidérons l'exemple des taux de change de versements réalisés par Coopération Suisse: {590,6; 550,3; 554,5; 565,2; 593,4} (Tableau 3) dans Kpenavoun Chogou (2020). La moyenne harmonique est $\widehat{m} = 570{,}2$ Francs FCFA/Franc Suisse. La moyenne harmonique corrigée pour le biais est ici $\widehat{m}_c = \left(1 - \widehat{m}^2\widehat{\sigma_o^2}\right)\widehat{m}$ et l'erreur type est $\widehat{\sigma}_{\widehat{m}} = \widehat{m}^2\widehat{\sigma}_o$ avec $\widehat{\sigma_o^2}$ la variance estimée pour l'inverse du taux de change: $\widehat{\sigma_o^2} = 7{,}55 \times 10^{-10}$. On obtient ainsi un taux de change moyen de $\widehat{m} = 570{,}1$ Francs FCFA/Franc Suisse avec une erreur type de $\widehat{\sigma}_{\widehat{m}} = 8{,}9$ Francs FCFA/Franc Suisse.

Plus généralement, muni d'une transformation monotone $\psi(\cdot)$ et de sa réciproque $\psi^{-1}(\cdot)$, le biais d'une moyenne quasi-arithmétique définie par $\widehat{m} = \psi^{-1}(T_o)$ avec $T_o = n^{-1}\sum_{i=1}^n \psi(x_i)$ est donné par

$$Biais_2(\widehat{m}) = \frac{h(T_o)}{2}\widehat{\sigma_o^2} \qquad \text{(approximation à l'ordre 2)}, \tag{4}$$

l'estimateur corrigé pour le biais est

$$\widehat{m}_c = \widehat{m} - \frac{h(T_o)}{2}\widehat{\sigma_o^2} \qquad \text{et} \tag{5}$$

l'erreur type de $\widehat{m}$ ou $\widehat{m}_c$ est estimée par

$$\sigma_{\widehat{m}} = \widehat{\sigma}_o|g(T_o)| \qquad \text{(approximation à l'ordre 1)} \tag{6}$$

où $\widehat{\sigma_o^2} = \frac{1}{n(n-1)}\sum_{i=1}^n (\psi(x_i) - T_o)^2$, et $g(\cdot)\cdot$ et $h(\cdot)$ sont respectivement les dérivées première et seconde de $\psi(\cdot)$. Une estimation plus précise (approximation à l'ordre 2) de la variance de $\widehat{m}$ peut être obtenue:

$$\sigma^2_{\widehat{m}_2} = [g(T_o)]^2\widehat{\sigma_o^2} + g(T_o)h(T_o)\widehat{\delta_o^3} + [h(T_o)]^2\left[\frac{1}{4}\widehat{\sigma^2_{T_o^2}} - T_o\left(\widehat{\delta_o^3} + T_o\widehat{\sigma_o^2}\right)\right] \quad \text{avec} \tag{7}$$

$$\widehat{\delta_o^3} = \frac{1}{n^2}\widehat{\delta_T^3} \tag{8}$$

$$\widehat{\sigma^2_{T_o^2}} = \frac{1}{n^3}\left[\widehat{\tau^4_T} + 4n\widehat{\mu_T\delta^3_T} + (2n-3)\widehat{\sigma^4_T} + 4n^2\widehat{\mu^2_T\sigma^2_T}\right] \quad (9)$$

$$\widehat{\delta^3_T} = \frac{n^2}{(n-1)(n-2)}R^3_T \quad (10)$$

$$\widehat{\sigma^4_T} = \frac{n^2(n^2-3n+3)}{n^3-5n^2+9n-3}\left[\frac{S^4_T}{n-1} - \frac{Q^4_T}{n^2-3n+3}\right] \quad (11)$$

$$\widehat{\tau^4_T} = \frac{1}{(n^2-3n+3)}\left[\frac{n^3}{n-1}Q^4_T - 3(n-3)\widehat{\sigma^4_T}\right] \quad (12)$$

$$\widehat{\mu_T\delta^3_T} = \frac{1}{n^2-8n-30}\left[n(n-2)T_o\widehat{\delta^3_T} - 3n(n+5)T_o^2\widehat{\sigma^2_T} - \frac{n^2-5n-15}{n}\widehat{\tau^4_T} + \frac{3(2n^2-15)}{n}\widehat{\sigma^4_T}\right] \quad (13)$$

$$\widehat{\mu^2_T\sigma^2_T} = T_o^2\widehat{\sigma^2_T} - \frac{\widehat{\tau^4_T} + (n-3)\widehat{\sigma^4_T}}{n^2} - \frac{\widehat{\mu_T\delta^3_T}}{n} \quad (14)$$

où $R^3_T = n^{-1}\sum_{i=1}^n [\psi(x_i) - T_o]^3$ et $Q^4_T = n^{-1}\sum_{i=1}^n [\psi(x_i) - T_o]^4$.

3.3 Intervalles de confiance autour des moyennes (quasi) arithmétiques

3.3.1 Cas de la moyenne arithmétique

Ici, la moyenne arithmétique $\widehat{m} = \bar{X}$ estime l'espérance mathématique m d'une population. Si la variance σ^2 de X est connue, un intervalle de confiance asymptotique (pour n grand) au niveau $100(1-\alpha)\%$ est

$$IC_{1-\alpha}(m) = \left[\bar{X} - \frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}}\right] \quad (15)$$

où $Z_{1-\frac{\alpha}{2}}$ est le percentile d'ordre $1-\frac{\alpha}{2}$ de la distribution normal standard. Si $\alpha = 5\%$, $Z_{97.5\%} \approx 1.95996$ généralement arrondi à $Z_{97.5\%} \approx 1.96$. Mais en pratique, σ^2 est souvent inconnue et la variance $\frac{\sigma^2}{n}$ de $\bar{X}$ est estimée par $\widehat{\sigma^2_o} = \frac{1}{n(n-1)}\sum_{i=1}^n (x_i - \bar{X})^2$. Dans ce cas,

$$IC_{1-\alpha}(m) = \left[\bar{X} - \widehat{\sigma}_o t^{n-1}_{1-\frac{\alpha}{2}}, \bar{X} + \widehat{\sigma}_o t^{n-1}_{1-\frac{\alpha}{2}}\right] \quad (16)$$

est un intervalle de confiance au niveau $100(1-\alpha)\%$, avec $t^{n-1}_{1-\frac{\alpha}{2}}$ le percentile d'ordre $1-\frac{\alpha}{2}$ de la distribution t de Student de $n-1$ degrés de liberté.

3.3.2 Cas des moyennes quasi-arithmétiques

Dans les cas de moyennes quasi-arithmétiques, un intervalle de confiance asymptotique au niveau $100(1-\alpha)\%$ autour de la moyenne m estimée par $\widehat{m} = \psi^{-1}(T_o)$ avec $T_o = n^{-1}\sum_{i=1}^n \psi(x_i)$ est

$$IC_{1-\alpha}(m) = \left[\psi^{-1}\left(T_o - \widehat{\sigma}_o t^{n-1}_{1-\frac{\alpha}{2}}\right),\ \psi^{-1}\left(T_o + \widehat{\sigma}_o t^{n-1}_{1-\frac{\alpha}{2}}\right)\right] \quad \text{si } \psi^{-1}(\cdot) \text{ est croissante et} \quad (17)$$

$$IC_{1-\alpha}(m) = \left[\psi^{-1}\left(T_o + \widehat{\sigma}_o t^{n-1}_{1-\frac{\alpha}{2}}\right),\ \psi^{-1}\left(T_o - \widehat{\sigma}_o t^{n-1}_{1-\frac{\alpha}{2}}\right)\right] \quad \text{si } \psi^{-1}(\cdot) \text{ est decroissante} \quad (18)$$

avec $\widehat{\sigma}_o = \sqrt{\widehat{\sigma_o^2}}$ (voir Tableau 1 ou les équations (6) et (7)).

Pour l'exemple des taux de change de versements réalisés par la Coopération Suisse, la taille de l'échantillon considéré est $n = 5$. Afin de construire un intervalle de confiance de niveau 95% autour de la moyenne harmonique de la population des taux, on détermine les statistiques $T_o = 0{,}0018$ et $\widehat{\sigma}_o = 2{,}7477 \times 10^{-5}$; et le quantile $t^4_{97{,}5\%} = 2{,}7764$. On obtient alors $IC_{95\%}(m) = [546{,}5; 596{,}2]$ Francs FCFA/Franc Suisse. Avec un niveau de confiance de 90%, on obtient avec $t^4_{95\%} = 2{,}1318$, $IC_{90\%}(m) = [551{,}8; 589{,}9]$ Francs FCFA/Franc Suisse.

Conclusion

Il convient de toujours faire ressortir la signification d'une moyenne lorsqu'on la demande où lors de son interprétation. Comme souligné par Kpenavoun Chogou (2020), il n'est nullement nécessaire de retenir le type de moyenne suivant la nature des données, quoiqu'une telle connaissance s'établit généralement grâce à l'usage répétitif de différents types de moyennes. Le plus important est l'adéquation entre l'interprétation connue d'un type de moyenne et l'objectif que vise le calcul de la moyenne. Tout professionnel disposant de la signification réelle d'une moyenne devrait pouvoir la calculer sans erreur. Pour exemple de significations précises (voir détails dans les exemples de Kpenavoun Chogou (2020)), la moyenne quadratique de la longueur du côté d'une parcelle carrée est *la longueur du côté de la parcelle de superficie moyenne*; la moyenne harmonique d'un taux de change est *le taux de conversion du montant moyen transféré en montant moyen reçu*; la moyenne géométrique du coefficient multiplicateur du salaire est *le coefficient correspondant à une variation relative constante du salaire.*

En claire, des erreurs d'interprétation sont commises dans le calcul des moyennes simplement parce que des erreurs sont commises dans la demande de ces moyennes. Enfin, il est important de toujours faire suivre une moyenne d'un indicateur de precision (erreur type) ou d'un intervalle de confiance.

References

PS Bullen. *Means and their inequalities*. Springer, Netherlands, 5 edition, 2013. ISBN 9789401703994, 940170399X.

M Caveing. Histoire des mathématiques de l'antiquité. *Revue de synthèse*, 4(4), 1998.

M de Carvalho. Mean, What do you Mean? *The American Statistician*, 70(3):270–274, 2016.

C Goldfarb, B et Pardoux. *Introduction à la méthode statistique: Manuel et exercices corrigés*. Dunod, Paris, 6 edition, 2011.

A Kolmogorov. *Sur la notion de la moyenne*, volume 12. Atti della Academia Nazionale dei Lincei, 6 edition, 1930.

S Kpenavoun Chogou. Calcul de la moyenne d'une variable quantitative: des erreurs sont commises, 2020. URL DOI:10.13140/RG.2.2.22049.56161/2. Working Paper N°01/2020/UAC/FSA/EESAC/LEPPA.

JL Marichal. *Fonctions d'agrégation pour la décision*, 2006. URL http://hdl.handle.net/10993/6891. Mathematiques appliquées, Université du Luxembourg.

R Palm. Utilisation du bootstrap pour les problèmes statistiques liés à l'estimation des paramètres. *Biotechnologie, agronomie, société et environnement*, 6(3), 2002.

P Prescott. Selection of Trimming Proportions for Robust Adaptive Trimmed Means. *Journal of the American Statistical Association*, 73(361):133–140, 1978.

U Ricci. Confronti tra medie. (italian). *Giorn Economisti e Rivista di Statistica*, 26(1):38–66, 1915.

J Rondeux. *Lexique des principaux termes dendrométriques*. Faculté des Sciences Agronomiques de l'Etat, Gembloux, 1 edition, 1978.

4 Annexe

4.1 Signification d'une moyenne géométrique

Cet appendice utilise l'exemple du coefficient multiplicatif du salaire pour dériver la signification d'une moyenne géométrique. Notons $S(t)$ le salaire à un instant t. Supposons que la variation relative du salaire est constante et égale à $\alpha \in]0;\infty[$, *i.e.* $\frac{\partial S(t)/\partial t}{S(t)} = \alpha$.

L'hypothèse de la variation relative constante conduit (après résolution de l'équation différentielle $\frac{\partial S(t)/\partial t}{S(t)} = \alpha$) à un salaire de la forme $S(t) = S_o \mathrm{e}^{\alpha t + \epsilon_t}$ où S_o est le salaire initial et ϵ_t est une réalisation au temps t d'une variable aléatoire quelconque ϵ d'espérance mathématique $E\{\epsilon\} = a$ avec a une constante réelle. Le coefficient multiplicatif $C(t) = \frac{S(t)}{S(t-1)}$ du salaire satisfait alors $C(t) = \mathrm{e}^{\alpha + e_t}$ avec $e_t = \epsilon_t - \epsilon_{t-1}$. Notons que $E\{e_t\} = a - a = 0$. En prenant le logarithme de $C(t)$ on a $\log C(t) = \alpha + e_t$ d'où $E\{\log C(t)\} = \alpha$. En d'autres termes, la moyenne géométrique $m = \exp\left(E\{\log C(t)\}\right)$ du coefficient $C(t)$ est l'exponentielle de la variation relative constante α: $m = \mathrm{e}^{\alpha}$. Ainsi, en ignorant le terme d'erreur e_t dans l'expression $C(t) = \mathrm{e}^{\alpha + e_t}$ du coefficient, la moyenne géométrique du coefficient multiplicatif est le coefficient correspondant à une variation relative constante.

4.2 Moments de la moyenne arithmétique

Considérons une variable aléatoire réelle X distribuée selon une loi quelconque $\mathcal{L}$ dont l'espérance mathématique, la variance, et les moments centrés d'ordres 3 et 4 sont $\mu = E\{X\}$, $\sigma^2 = E\{(X-\mu)^2\}$, $\delta^3 = E\{(X-\mu)^3\}$ et $\tau^4 = E\{(X-\mu)^4\}$. Les coefficients d'asymétrie et d'aplatissement (kurtose) de X sont respectivement définis par $\beta = \frac{\delta^3}{\sigma^3}$ et $\gamma = \frac{\tau^4}{\sigma^4} - 3$. Les moments non centrés d'ordres 2 à 4 sont par suite $E\{X^2\} = \sigma^2 + \mu^2$, $E\{X^3\} = \delta^3 + \mu(3\sigma^2 + \mu^2)$ et $E\{X^4\} = \tau^4 + 4\mu\delta^3 + \mu^2(6\sigma^2 + \mu^2)$.

Considérons ensuite une série statistique $\mathcal{X} = \{x_1, x_2, \cdots x_n\}$ provenant d'un échantillon aléatoire et simple, *i.e.* les x_i $(i = 1, 2, \cdots, n)$ sont identiquement et indépendamment distribués suivant la même loi $\mathcal{L}$. La moyenne moyenne arithmétique $\bar{X}$ est définie pour la série $\mathcal{X}$ par $\bar{X} = n^{-1} \sum_{i=1} x_i$. Les quatre premiers moments (centrés ou non) de la moyenne arithmétique $\bar{X}$ sont déterminés dans cet appendice.

4.2.1 Espérance mathématique de la moyenne arithmétique

L'espérance mathématique $\mu_o = E\left\{\bar{X}\right\}$ de la moyenne arithmétique est trouvée a travers

$$\begin{aligned} \mu_o &= \frac{1}{n} E\left\{\sum_{i=1}^{n} x_i\right\} = \frac{1}{n}\sum_{i=1}^{n} E\left\{x_i\right\} = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu \\ \mu_o &= \mu. \end{aligned} \tag{19}$$

4.2.2 Moment d'ordre 2 de la moyenne arithmétique

A partir du développement $\left(\sum_{i=1}^{n} x_i\right)^2 = \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n}\sum_{j\neq i} x_i x_j$ et des identités $E\left\{x_i^2\right\} = \sigma^2 + \mu^2$ et $E\left\{x_i x_j\right\} = \mu^2$ pour $i \neq j$, l'espérance du carré de la moyenne arithmétique est

$$\begin{aligned} E\left\{\left(\bar{X}\right)^2\right\} &= \frac{1}{n^2}\left[\sum_{i=1}^{n} E\left\{x_i^2\right\} + \sum_{i=1}^{n}\sum_{j\neq i} E\left\{x_i x_j\right\}\right] \\ &= \frac{1}{n^2}\left[n(\sigma^2+\mu^2) + n(n-1)\mu^2\right] \\ E\left\{\left(\bar{X}\right)^2\right\} &= \frac{\sigma^2}{n} + \mu^2. \end{aligned} \tag{20}$$

On déduit des équations (19) et (20), l'expression de la variance $\sigma_o^2 = E\left\{\left(\bar{X}\right)^2\right\} - \left[E\left\{\bar{X}\right\}\right]^2$ de $\bar{X}$,

$$\sigma_o^2 = \frac{\sigma^2}{n}. \tag{21}$$

4.2.3 Moment d'ordre 3 de la moyenne arithmétique

A partir du développement $\left(\sum_{i=1}^{n} x_i\right)^3 = \sum_{i=1}^{n} x_i^3 + 3\sum_{i=1}^{n}\sum_{j\neq i} x_i x_j^2 + \sum_{i=1}^{n}\sum_{j\neq i}\sum_{k\neq i,j} x_i x_j x_k$ et des identités $E\{x_i^3\} = \delta^3 + \mu(3\sigma^2+\mu^2)$, $E\left\{x_i x_j^2\right\} = \mu(\sigma^2+\mu^2)$ et $E\left\{x_i x_j x_k\right\} = \mu^3$ pour $i \neq j \neq k$, l'espérance cubique de la moyenne arithmétique est

$$\begin{aligned} E\left\{\left(\bar{X}\right)^3\right\} &= \frac{1}{n^3}\left[\sum_{i=1}^{n} E\left\{x_i^3\right\} + 3\sum_{i=1}^{n}\sum_{j\neq i} E\left\{x_i x_j^2\right\} + \sum_{i=1}^{n}\sum_{j\neq j}\sum_{k\neq i,j} E\left\{x_i x_j x_k\right\}\right] \\ &= \frac{1}{n^3}\left[n\left[\delta^3 + \mu(3\sigma^2+\mu^2)\right] + 3n(n-1)\mu(\sigma^2+\mu^2) + n(n-1)(n-2)\mu^3\right] \\ E\left\{\left(\bar{X}\right)^3\right\} &= \frac{1}{n^2}\left[\delta^3 + \mu\left(3n\sigma^2 + n^2\mu^2\right)\right]. \end{aligned} \tag{22}$$

On en déduit le moment centré d'ordre 3 de $\bar{X}$:

$$\begin{aligned} \delta_o^3 &= E\left\{\left(\bar{X}-\mu_o\right)^3\right\} = E\left\{\left(\bar{X}\right)^3\right\} - \mu_o(3\sigma_o^2 + \mu_o^2) \\ &= \frac{1}{n^2}\left[\delta^3 - \mu\left(3n\sigma^2 + n^2\mu^2\right)\right] - \mu(3\frac{\sigma^2}{n} + \mu^2) \end{aligned}$$

$$\delta_o^3 = \frac{\delta^3}{n^2} \tag{23}$$

ainsi que le coefficient d'asymétrie de $\bar{X}$:

$$\beta_o = \frac{E\left\{\left(\bar{X}-\mu_o\right)^3\right\}}{\sigma_o^3} = \frac{n^{3/2}}{\sigma^3}\frac{\delta^3}{n^2}$$

$$\beta_o = \frac{\delta^3}{\sigma^3\sqrt{n}} = \frac{\beta}{\sqrt{n}}. \tag{24}$$

4.2.4 Moment d'ordre 4 de la moyenne arithmétique

A partir du développement

$$\left(\sum_{i=1}^n x_i\right)^4 = \sum_{i=1}^n x_i^4 + 4\sum_{i=1}^n\sum_{j\neq i} x_i x_j^3 + 3\sum_{i=1}^n\sum_{j\neq i} x_i^2 x_j^2 + 6\sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j} x_i x_j x_k^2 + \sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} x_i x_j x_k x_l$$

et des expressions des moments $E\{x_i^4\} = \tau^4+4\mu\delta^3+\mu^2(6\sigma^2+\mu^2)$, $E\{x_i x_j^3\} = \mu\left[\delta^3+\mu(3\sigma^2+\mu^2)\right]$, $E\left\{x_i^2x_j^2\right\} = (\sigma^2+\mu^2)^2$, $E\left\{x_i x_j x_k^2\right\} = \mu^2(\sigma^2+\mu^2)$ et $E\left\{x_i x_j x_k x_l\right\} = \mu^4$ pour $i\neq j\neq k\neq l$,

l'espérance quartique de la moyenne arithmétique est

$$E\left\{\left(\bar{X}\right)^4\right\} = \frac{1}{n^4}\left[\sum_{i=1}^n E\left\{x_i^4\right\} + 4\sum_{i=1}^n\sum_{j\neq j} E\left\{x_i x_j^3\right\} + 3\sum_{i=1}^n\sum_{j\neq j} E\left\{x_i^2x_j^2\right\} + 6\sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j} E\left\{x_i x_j x_k^2\right\} + \sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j}\sum_{l\neq i,j,k} E\left\{x_i x_j x_k x_l\right\}\right]$$

$$= \frac{1}{n^3}\left[\tau^4 + 4\mu\delta^3 + \mu^2(6\sigma^2+\mu^2) + 4(n-1)\mu\left[\delta^3+\mu(3\sigma^2+\mu^2)\right] + 3(n-1)(\sigma^2+\mu^2)^2 + 6(n-1)(n-2)\mu^2(\sigma^2+\mu^2) + (n-1)(n-2)(n-3)\mu^4\right]$$

$$E\left\{\left(\bar{X}\right)^4\right\} = \frac{1}{n^3}\left[\tau^4 + 4n\mu\delta^3 + 3(n-1)\sigma^4 + 6n^2\mu^2\sigma^2 + n^3\mu^4\right]. \tag{25}$$

On en déduit le moment centré d'ordre 4 de $\bar{X}$:

$$\tau_o^4 = E\left\{\left(\bar{X}-\mu_o\right)^4\right\} = E\left\{\left(\bar{X}\right)^4\right\} - 4\mu_o\delta_o^3 - \mu_o^2(6\sigma_o^2+\mu_o^2)$$

$$= \frac{1}{n^3}\left[\tau^4 + 4n\mu\delta^3 + 3(n-1)\sigma^4 + 6n^2\mu^2\sigma^2 + n^3\mu^4 - 4n^3\mu\frac{\delta^3}{n^2} - 6n^3\mu^2\frac{\sigma^2}{n} - n^3\mu^4\right]$$

$$\tau_o^4 = \frac{1}{n^3}\left[\tau^4 + 3(n-1)\sigma^4\right] \tag{26}$$

ainsi que le coefficient d'aplatissement de $\bar{X}$:

$$\begin{aligned}\gamma_o &= \frac{E\left\{\left(\bar{X}-\mu_o\right)^4\right\}}{\sigma_o^4} - 3 = \frac{n^2}{\sigma^4}\frac{\tau^4+3(n-1)\sigma^4}{n^3} - 3 \\ &= \frac{1}{n}\left(\frac{\tau^4}{\sigma^4} - 3 + 3n\right) - 3 \\ \gamma_o &= \frac{\gamma}{n}. \end{aligned} \tag{27}$$

La variance du carré de la moyenne arithmétique est donnée par

$$\begin{aligned}\sigma^2_{\bar{X}^2} &= E\left\{\left(\bar{X}\right)^4\right\} - \left[E\left\{\left(\bar{X}\right)^2\right\}\right]^2 \\ \sigma^2_{\bar{X}^2} &= \frac{1}{n^3}\left[\tau^4 + 4n\mu\delta^3 + (2n-3)\sigma^4 + 4n^2\mu^2\sigma^2\right]. \end{aligned} \tag{28}$$

4.3 Estimation de moments centrés de la moyenne arithmétique

L'équation (19) indique que $\bar{X}$ et X ont la même espérance mathématique. Il en result que $\bar{X}$ est un estimateur non biaisé de l'espérance de X,

$$\widehat{\mu} = \bar{X}. \tag{29}$$

Cette section détermine les espérances mathématiques des moments centrés d'ordre 2, 3 et 4 de X pour en déduire des estimateurs non biaisés des moments centrés σ^2, δ^3 et τ^4 ainsi que des produits μ^2, μ^3, $\mu\sigma^2$, μ^4, $\mu^2\sigma^2$, $\mu\delta^3$ et σ^4.

4.3.1 Espérance mathématique de la variance d'un échantillon

La variance d'une série $\mathcal{X}$ est définie par $S_X^2 = n^{-1}\sum_{i=1}^n \left(x_i - \bar{X}\right)^2$. En utilisant la forme développée $S_X^2 = n^{-1}\sum_{i=1}^n x_i^2 - \left(\bar{X}\right)^2$ de S_X^2, l'identité $E\left\{x_i^2\right\} = \sigma^2 + \mu^2$ et le moment d'ordre 2 de la moyenne arithmétique donné en (20), l'espérance mathématique de S_X^2 est donnée par

$$\begin{aligned}E\left\{S_X^2\right\} &= \frac{1}{n}\sum_{i=1}^n E\left\{x_i^2\right\} - E\left\{\left(\bar{X}\right)^2\right\} \\ &= \left(\sigma^2+\mu^2\right) - \frac{1}{n}\left(\sigma^2 + n\mu^2\right) \\ E\left\{S_X^2\right\} &= \frac{n-1}{n}\sigma^2. \end{aligned} \tag{30}$$

Il en ressort que la variance d'un échantillon est un estimateur consistent ($E\left\{S_X^2\right\} \to \sigma^2$ quand $n \to \infty$), mais biaisé de la variance de la population mère: $E\left\{S_X^2\right\}$ sous estime σ^2. On déduit de (30) un estimateur non biaisé de σ^2:

$$\widehat{\sigma^2} = \frac{n}{n-1}S_X^2$$

$$\widehat{\sigma^2} \;=\; \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^2. \tag{31}$$

De la relation $E\left\{\left(\bar{X}\right)^2\right\}=\frac{\sigma^2}{n}+\mu^2$ (20), on déduit aussi qu'un estimateur non biaisé du carré μ^2 de l'espérance mathématique de X est

$$\widehat{\mu^2} \;=\; \left(\bar{X}\right)^2-\frac{S_X^2}{n-1}. \tag{32}$$

4.3.2 Espérance mathématique du moment centré d'ordre 3

Le moment centré cubique d'une série $\mathcal{X}$ est définie par $R_X^3=n^{-1}\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^3$, ou de façon équivalente par l'expression plus succincte $R_X^3=\frac{1}{n}\sum_{i=1}^{n}x_i^3-n\bar{X}\left(3S_X^2+\left(\bar{X}\right)^2\right)$. En considérant le développement $R_X^3=\frac{1}{n}\sum_{i=1}^{n}x_i^3-\frac{3}{n^2}\left(\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_i^2+2\left(\bar{X}\right)^3$ de R_X^3, l'espérance de $\left(\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_i^2$ donnée par,

$$\begin{aligned}
E\left\{\left(\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_i^2\right\} &= \sum_{i=1}^{n}E\left\{x_i^3\right\}+\sum_{i=1}^{n}\sum_{j\neq i}E\left\{x_ix_j^2\right\}\\
&= n[\delta^3+\mu(3\sigma^2+\mu^2)]+n(n-1)\mu(\sigma^2+\mu^2)\\
&= n[\delta^3+(n+2)\mu\sigma^2+n\mu^3]
\end{aligned} \tag{33}$$

et le moment d'ordre 3 de la moyenne arithmétique donné en (22), l'espérance de R_X^3 est donnée par

$$\begin{aligned}
E\left\{R_X^3\right\} &= \frac{1}{n}\sum_{i=1}^{n}E\left\{x_i^3\right\}-\frac{3}{n^2}E\left\{\left(\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_i^2\right\}+2E\left\{\left(\bar{X}\right)^3\right\}\\
&= \delta^3+\mu(3\sigma^2+\mu^2)-\frac{3}{n}\left[\delta^3+(n+2)\mu\sigma^2+n\mu^3\right]+\frac{2}{n^2}\left[\delta^3+3n\mu\sigma^2+n^2\mu^3\right]\\
E\left\{R_X^3\right\} &= \frac{(n-1)(n-2)}{n^2}\delta^3.
\end{aligned} \tag{34}$$

Il en ressort que R_X^3 est un estimateur consistent mais biaisé du moment centré d'ordre 3 (δ^3) de la population mère. On déduit de (34) un estimateur non biaisé de δ^3:

$$\begin{aligned}
\widehat{\delta^3} &= \frac{n^2}{(n-1)(n-2)}R_X^3\\
\widehat{\delta^3} &= \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(x_i-\bar{X}\right)^3.
\end{aligned} \tag{35}$$

De plus, à partir de $E\left\{\bar{X}S_X^2\right\}=\frac{1}{n^2}E\left\{\left(\sum_{i=1}^{n}x_i\right)\sum_{i=1}^{n}x_i^2\right\}-E\left\{\left(\bar{X}\right)^3\right\}$ réduit en utilisant (33) à

$$E\left\{\bar{X}S_X^2\right\}=\frac{n-1}{n^2}\left[\delta^3+n\mu\sigma^2\right] \tag{36}$$

on obtient un estimateur non biaisé du produit $\mu\sigma^2$:

$$\begin{aligned}\widehat{\mu\sigma^2} &= \widehat{\mu}\widehat{\sigma^2} - \frac{\widehat{\delta^3}}{n} \\ \widehat{\mu\sigma^2} &= \frac{n}{n-1}\bar{X}S_X^2 - \frac{n}{(n-1)(n-2)}R_X^3. \end{aligned} \tag{37}$$

De la relation $E\left\{\left(\bar{X}\right)^3\right\} = \frac{\delta^3}{n^2} + \frac{3\mu\sigma^2}{n} + \mu^3$ (22), on déduit qu'un estimateur non biaisé du cube μ^3 de l'espérance mathématique de X est

$$\begin{aligned}\widehat{\mu^3} &= \widehat{\mu}^3 - \frac{3\widehat{\mu\sigma^2}}{n} - \frac{\widehat{\delta^3}}{n^2} \\ \widehat{\mu^3} &= \left(\bar{X}\right)^3 + \frac{2}{(n-1)(n-2)}R_X^3 - \frac{3}{n-1}\bar{X}S_X^2. \end{aligned} \tag{38}$$

4.3.3 Espérance mathématique du moment centré d'ordre 4

Le moment centré quartique d'une série $\mathcal{X}$ est définie par $Q_X^4 = n^{-1}\sum_{i=1}^n\left(x_i - \bar{X}\right)^4$. En utilisant la forme développée $Q_X^4 = \frac{1}{n}\sum_{i=1}^n x_i^4 - \frac{4}{n^2}\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i^3 + \frac{6}{n^3}\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2 - 3\left(\bar{X}\right)^4$ de Q_X^4, l'espérance de $\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i^3$ donnée par,

$$\begin{aligned}E\left\{\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i^3\right\} &= \sum_{i=1}^n E\left\{x_i^4\right\} + \sum_{i=1}^n\sum_{j\neq i} E\left\{x_i x_j^3\right\} \\ &= n\left[\tau^4 + 4\mu\delta^3 + 6\mu^2\sigma^2 + \mu^4\right] + n(n-1)\mu\left[\delta^3 + 3\mu\sigma^2 + \mu^3\right] \\ &= n\left[\tau^4 + (n+3)\mu\delta^3 + 3(n+1)\mu^2\sigma^2 + n\mu^4\right], \end{aligned} \tag{39}$$

l'espérance de $\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2$ donnée par,

$$\begin{aligned}E\left\{\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2\right\} &= \sum_{i=1}^n E\left\{x_i^4\right\} + \sum_{i=1}^n\sum_{j\neq i} E\left\{x_i^2 x_j^2\right\} + 2\sum_{i=1}^n\sum_{j\neq i} E\left\{x_i x_j^3\right\} + \sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j} E\left\{x_i x_j x_k^2\right\} \\ &= n\left(\tau^4 + 4\mu\delta^3 + 6\mu^2\sigma^2 + \mu^4\right) + n(n-1)(\sigma^2+\mu^2)^2 \\ &\quad + 2n(n-1)\mu\left(\delta^3 + 3\mu\sigma^2 + \mu^3\right) + n(n-1)(n-2)\mu^2\left(\sigma^2+\mu^2\right) \\ &= n\left[\tau^4 + (n-1)\sigma^4 + 2(n+1)\mu\delta^3 + n(n+5)\mu^2\sigma^2 + n^2\mu^4\right], \end{aligned} \tag{40}$$

et le moment d'ordre 4 de la moyenne arithmétique donné en (25), l'espérance de Q_X^4 est donnée par

$$\begin{aligned}E\left\{Q_X^4\right\} &= \frac{1}{n}\sum_{i=1}^n E\left\{x_i^4\right\} - \frac{4}{n^2}E\left\{\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i^3\right\} + \frac{6}{n^3}E\left\{\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2\right\} - 3E\left\{\left(\bar{X}\right)^4\right\} \\ E\left\{Q_X^4\right\} &= \frac{(n-1)(n^2-3n+3)}{n^3}\tau^4 + \frac{3(n-1)(2n-3)}{n^3}\sigma^4. \end{aligned} \tag{41}$$

Il en ressort que Q_X^4 est un estimateur consistent mais biaisé du moment centré d'ordre 4 (τ^4) de la population mère. Afin de pouvoir déduire de (41) un estimateur non biaisé de τ^4, il est nécessaire de trouver en premier lieu un estimateur non biaisé de σ^4. Pour ce faire, on détermine d'abord l'espérance mathématique de $S_X^4 = \left[S_X^2\right]^2 = \frac{1}{n^2}\left(\sum_{i=1}^n x_i^2\right)\sum_{i=1}^n x_i^2 - \frac{2}{n^3}\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n x_i\right)^2 + \left(\bar{X}\right)^4$. On obtient à partir des développements de $\left(\sum_{i=1}^n x_i^2\right)\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i^4 + \sum_{i=1}^n\sum_{j\neq i} x_i^2 x_j^2$ ainsi que de $\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n x_i^4 + \sum_{i=1}^n\sum_{j\neq i} x_i^2 x_j^2 + 2\sum_{i=1}^n\sum_{j\neq i} x_i x_j^3 + \sum_{i=1}^n\sum_{j\neq i}\sum_{k\neq i,j} x_i x_j x_k^2$,

$$E\left\{\left[S_X^2\right]^2\right\} = \frac{n-1}{n^3}\left[(n-1)\tau^4 + (n^2-2n+3)\sigma^4\right]. \tag{42}$$

En remplaçant $E\left\{Q_X^4\right\}$ par Q_X^4 dans (41) et $E\left\{S_X^4\right\}$ par S_X^4 dans (42) et en les résolvant conjointement pour σ^4 et τ^4, on obtient les estimateurs non biaisés:

$$\widehat{\tau^4} = \frac{1}{(n^2-3n+3)}\left[\frac{n^3}{n-1}Q_X^4 - 3(n-3)\widehat{\sigma^4}\right] \tag{43}$$

$$\widehat{\sigma^4} = \frac{n^2(n^2-3n+3)}{n^3-5n^2+9n-3}\left[\frac{S_X^4}{n-1} - \frac{Q_X^4}{n^2-3n+3}\right]. \tag{44}$$

A partir des expressions (25), (39) et (40), on obtient les espérances

$$\begin{aligned} E\left\{\left(\bar{X}\right)^2 S_X^2\right\} &= \frac{1}{n^3}E\left\{\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2\right\} - E\left\{\left(\bar{X}\right)^4\right\} \\ &= \frac{n-1}{n^3}\left[\tau^4 + (n-3)\sigma^4 + 2n\mu\delta^3 + n^2\mu^2\sigma^2\right] \quad \text{et} \end{aligned} \tag{45}$$

$$\begin{aligned} E\left\{\bar{X}R_X^3\right\} &= \frac{1}{n^2}E\left\{\left(\sum_{i=1}^n x_i\right)\sum_{i=1}^n x_i^3\right\} - \frac{3}{n^3}E\left\{\left(\sum_{i=1}^n x_i\right)^2\sum_{i=1}^n x_i^2\right\} + 2E\left\{\left(\bar{X}\right)^4\right\} \\ &= \frac{n-1}{n^3}\left[(n-2)\tau^4 - 3(n-2)\sigma^4 + n(n-2)\mu\delta^3 + 3n(n+5)\mu^2\sigma^2\right]. \end{aligned} \tag{46}$$

En remplaçant $E\left\{\left(\bar{X}\right)^2 S_X^2\right\}$ par $\left(\bar{X}\right)^2 S_X^2$ dans (45) et $E\left\{\bar{X}R_X^3\right\}$ par $\bar{X}R_X^3$ dans (46) avant de les resoudre conjointement pour les produits $\mu\delta^3$ et $\mu^2\sigma^2$, on obtient les estimateurs non biaisés:

$$\widehat{\mu^2\sigma^2} = \widehat{\mu}^2\widehat{\sigma^2} - \frac{\widehat{\tau^4} + (n-3)\widehat{\sigma^4}}{n^2} - \frac{\widehat{\mu\delta^3}}{n} \tag{47}$$

$$\widehat{\mu\delta^3} = \frac{1}{n^2-8n-30}\left[n(n-2)\widehat{\mu}\widehat{\delta^3} - 3n(n+5)\widehat{\mu}^2\widehat{\sigma^2} - \frac{n^2-5n-15}{n}\widehat{\tau^4} + \frac{3(2n^2-15)}{n}\widehat{\sigma^4}\right] \tag{48}$$

A partir du moment d'ordre 4 de $\bar{X}$ (25), on obtient un estimateur non biaisé de la puissance 4 de l'espérance μ

$$\widehat{\mu^4} = \bar{X}^4 - \frac{1}{n^3}\left[\widehat{\tau^4} + 4n\widehat{\mu\delta^3} + 3(n-1)\widehat{\sigma^4} + 6n^2\widehat{\mu^2\sigma^2}\right]. \tag{49}$$

4.4 Inférence sur une valeur moyenne

Cette section décrit comment obtenir des approximations du biais, de l'erreur type (mesure de précision) et un intervalle de confiance sur une estimation $\widehat{m}$ (3) de la valeur moyenne m (1) d'une variable aléatoire réelle X, à partir d'un échantillon $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$. Seules les moyennes quasi-arithmétiques sont considérées. En d'autres termes, seules les moyennes basées sur l'espérance mathématique $E[\cdot]$ comme mesure basique de tendance centrale sont traitées. On note dans la suite $\mu = E\{X\}$ et $\sigma^2 = E\{(X-\mu)^2\}$ l'espérance mathématique et la variance de X. Les moments centrés d'ordres 3 et 4 de X sont notés $\delta^3 = E\{(X-\mu)^3\}$ et $\tau^4 = E\{(X-\mu)^4\}$.

Muni d'une transformation monotone $\psi(\cdot)$ et de sa réciproque $\psi^{-1}(\cdot)$, notons T la transformée par $\psi^{-1}(\cdot)$ de X. La moyenne est ici définie par $m = \psi^{-1}(\mu_T)$ avec $\mu_T = E[T]$ et la moyenne estimée est donnée par $\widehat{m} = \psi^{-1}(T_o)$ avec $T_o = n^{-1}\sum_{i=1}^{n}\psi(x_i)$, la moyenne arithmétique de la série des valeurs observées de T: $\mathcal{T} = \{t_1, t_2, \cdots, t_n\}$, $t_i = \psi(x_i)$. La moyenne et la variance de T_o sont respectivement notées μ_o et σ_o^2 alors que ses moments centrés d'ordres 3 et 4 sont notés δ_o^3 et τ_o^4.

4.4.1 Calcul du biais d'une moyenne estimée

Le biais d'un estimateur est l'écart attendu entre une valeur estimé $\widehat{m}$ et de la vraie valeur m:

$$Biais(\widehat{m}) = E\{\widehat{m}\} - m. \tag{50}$$

Un estimateur est dit *non biaisé* si le biais associé est nul. C'est la cas de la moyenne arithmétique $\bar{X}$ comme estimateur de l'espérance mathématique μ: $Biais(\bar{X}) = E\{\bar{X}\} - \mu = 0$. On parle de *sous-estimation* si le biais est négatif et de *sur-estimation* si le biais est positif.

Une approximation du biais peut être dérivée à partir d'une approximation de la transformation $\psi^{-1}(\cdot)$ par son développement limité (série de Taylor) autour de l'espérance mathématique $\mu_o = \mu_T$ de T_o:

$$\psi^{-1}(T_o) = \psi^{-1}(\mu_o) + \sum_{k=1}^{\infty}\frac{\psi^{(-k)}(\mu_o)}{k!}(T_o - \mu_o)^k \tag{51}$$

où $\psi^{(-k)}(t) = \frac{d^k\psi^{-1}(t)}{dt^k}$ est la dérivée d'ordre k de $\psi^{-1}(\cdot)$. En notant $g(\cdot)$ la dérivée première de

$\psi^{-1}(\cdot)$, *i.e.* $g(t) = \frac{d\psi^{-1}(t)}{dt}$, le développement limité à l'ordre 1 de $\psi^{-1}(\cdot)$ est donnée par

$$\psi^{-1}(T_o) \approx \psi^{-1}(\mu_o) + g(\mu_o)(T_o - \mu_o). \tag{52}$$

Il en découle que $E\{\widehat{m}\} = E\{\psi^{-1}(T_o)\} \approx \psi^{-1}(\mu_o) + g(\mu_T)(E\{T_o\} - \mu_o) = \psi^{-1}(\mu_o)$ car $E\{T_o\} = \mu_o$. En remplaçant m par son expression $m = \psi^{-1}(\mu_T) = \psi^{-1}(\mu_o)$, l'approximation à l'ordre 1 du biais est:

$$Biais_1(\widehat{m}) = E\{\widehat{m}\} - m = \psi^{-1}(\mu_o) - \psi^{-1}(\mu_o) = 0. \tag{53}$$

Cette expression du biais est exacte si $\psi^{-1}(\cdot)$ est linéaire. En notant $h(\cdot)$ la dérivé seconde de $\psi^{-1}(\cdot)$, *i.e.* $h(t) = \frac{dg(t)}{dt}$, l'approximation d'ordre 2 de $\psi^{-1}(\cdot)$ est

$$\psi^{-1}(T_o) \approx \psi^{-1}(\mu_o) + g(\mu_o)(T_o - \mu_o) + \frac{h(\mu_o)}{2}(T_o - \mu_o)^2. \tag{54}$$

On a alors $E\{\widehat{m}\} = E\{\psi(T_o)\} \approx \psi(\mu_o) + \frac{h(\mu_o)}{2}\sigma_o^2$. Il en résulte que

$$Biais_2(\widehat{m}) = \frac{h(\mu_o)}{2}\sigma_o^2 \tag{55}$$

est l'approximation à l'ordre 2 du biais de $\widehat{m}$. Cette expression du biais est exacte si $\psi^{-1}(\cdot)$ est une fonction quadratique. Si $\mu_o = \mu_T$ et $\sigma_o^2 = n^{-1}\sigma_T^2$ ne sont pas disponibles, ils peuvent être remplacés dans (55) par leurs estimateurs non biaisés, respectivement $\widehat{\mu}_o = T_o$ et

$$\widehat{\sigma_o^2} = \frac{1}{n(n-1)}\sum_{i=1}^{n}[\psi(x_i) - T_o]^2. \tag{56}$$

Aux ordres 3 et 4, les biais de $\widehat{m}$ sont respectivement estimés par

$$Biais_3(\widehat{m}) = \frac{h(\mu_o)}{2}\sigma_o^2 + \frac{\psi^{(-3)}(\mu_o)}{6}\delta_o^3 \tag{57}$$

$$Biais_4(\widehat{m}) = \frac{h(\mu_o)}{2}\sigma_o^2 + \frac{\psi^{(-3)}(\mu_o)}{6}\delta_o^3 + \frac{\psi^{(-4)}(\mu_o)}{24}\tau_o^4 \tag{58}$$

avec $\psi^{(-3)}(\cdot)$ et $\psi^{(-4)}(\cdot)$ les dérivées d'ordre 3 et 4 de $\psi^{-1}(\cdot)$.

7.4.1.1 Cas de la moyenne arithmétique

Rappelons qu'ici $m = \mu$ et $\widehat{m} = \bar{X}$. Étant donné que $E\{\bar{X}\} = \mu$, le biais est exactement égale à $Biais(\bar{X}) = E\{\bar{X}\} - \mu = 0$, *i.e.* $\widehat{m}$ est non biaisé.

7.4.1.2 Cas de la moyenne quadratique

La moyenne est définie comme $m = \sqrt{E\{X^2\}}$ et est estimée par $\widehat{m} = \sqrt{T_o}$ avec $T_o = n^{-1}\sum_{i=1}^{n} x_i^2$. Notons que $m = \sqrt{\mu^2 + \sigma^2}$. La transformation $\psi(\cdot)$ est $\psi(t) = t^{1/2}$ et sa réciproque est $\psi^{-1}(x) = x^2$. On en déduit que $g(t) = \frac{1}{2}t^{-1/2}$ et $h(t) = -\frac{1}{4}t^{-3/2}$.

L'espérance mathématique de T_o est $\mu_o = \mu_T = \mu^2 + \sigma^2$. La variance de T_o est $\sigma_o^2 = E\{T_o^2\} - \mu_o^2$.

En utilisant $\left[\sum_{i=1}^n x_i^2\right]^2 = \sum_{i=1}^n x_i^4 + 2\sum_{i=1}^n \sum_{j\neq i} x_i^2 x_j^2$, $E\{x_i^4\} = \tau^4 + 4\mu\delta^3 + 6\mu^2\sigma^2 + \mu^4$ et $E\{x_i^2 x_j^2\} = (\mu^2+\sigma^2)^2$ pour $i \neq j$, on obtient $E\{T_o^2\} = n^{-1}\left[\tau^4 + 4\mu\delta^3 + (n-1)\sigma^4 + 2(n+2)\mu^2\sigma^2 + n\mu^4\right]$. La variance de T_o est ainsi $\sigma_o^2 = n^{-1}\left[\tau^4 + 4\mu\delta^3 + 4\mu^2\sigma^2 - \sigma^4\right]$. Par conséquent, l'approximation d'ordre 2 du biais de $\widehat{m}$ est:

$$Biais_2(\widehat{m}) = -\frac{\sigma_o^2}{8n(\mu^2+\sigma^2)^{3/2}} \quad \text{avec} \tag{59}$$

$$\sigma_o^2 = \frac{1}{n}\left(\tau^4 + 4\mu\delta^3 + 4\mu^2\sigma^2 - \sigma^4\right). \tag{60}$$

Il apparaît alors que la moyenne quadratique estimée $\widehat{m} = \sqrt{n^{-1}\sum_{i=1}^n x_i^2}$ est un estimateur biaisé de la moyenne quadratique de la population: $\widehat{m}$ sous-estime m (à moins que $\sigma^4 = \tau^4 + 4\mu\delta^3 + 4\mu^2\sigma^2$). En générale, les quantités μ, σ^2, δ^3 et τ^4 ne sont pas connues. On peut alors remplacer μ_o et σ_o^2 dans (55) par leurs estimateurs non biaisés respectifs $\widehat{\mu}_o = T_o$ et

$$\widehat{\sigma_o^2} = \frac{1}{n(n-1)}\sum_{i=1}^n (x_i^2 - T_o)^2. \tag{61}$$

Le biais approximatif est donné par

$$\widehat{Biais}_2(\widehat{m}) = -\frac{\widehat{\sigma_o^2}}{8T_o\sqrt{T_o}}. \tag{62}$$

Un estimateur moins biaisé de m (corrigé pour le biais) est par suite donnée par

$$\widehat{m}_c = \sqrt{T_o} + \frac{\widehat{\sigma_o^2}}{8T_o\sqrt{T_o}}. \tag{63}$$

7.4.1.3 Cas de la moyenne géométrique

La moyenne géométrique $m = \exp(E\{\log(X)\})$ est estimée par $\widehat{m} = \exp(T_o)$ avec $T_o = n^{-1}\sum_{i=1}^n \log(x_i)$. Notons que la transformation $\psi(\cdot)$ et ses deux premières dérivées sont $\psi(t) = g(t) = h(t) = \exp(t)$ et sa réciproque est $\psi^{-1}(x) = \log(x)$. Les quantités μ_o et σ_o^2 ne sont en générale pas disponibles et sont donc à remplacées par respectivement $\widehat{\mu}_o = T_o$ et

$$\widehat{\sigma_o^2} = \frac{1}{n(n-1)}\sum_{i=1}^n \left[\log(x_i) - T_o\right]^2. \tag{64}$$

Le bais estimé est donc donné par

$$\widehat{Biais}_2(\widehat{m}) = \frac{1}{2}\exp(T_o)\widehat{\sigma_o^2}. \tag{65}$$

Il apparaît alors que la moyenne géométrique estimée $\widehat{m} = \exp\left(n^{-1}\sum_{i=1}^n \log(x_i)\right)$ est un estimateur biaisé de la moyenne géométrique de la population: $\widehat{m}$ sur-estime m. Un estimateur moins biaisé de m est par suite

$$\widehat{m}_c = \left(1 - \widehat{\sigma_o^2}/2\right)\exp(T_o). \tag{66}$$

7.4.1.4 Cas de la moyenne harmonique

La moyenne harmonique $m = \left[E\{X^{-1}\}\right]^{-1}$ est estimée par $\widehat{m} = T_o^{-1}$ avec $T_o = n^{-1}\sum_{i=1}^{n} x_i^{-1}$. La transformation $\psi(\cdot)$ et sa réciproque sont $\psi(t) = \psi^{-1}(t) = t^{-1}$ et ses deux premières dérivées sont $g(t) = -t^{-2}$ et $h(t) = 2t^{-3}$. Les quantités μ_o et σ_o^2 ne sont en générale pas disponibles et sont donc à remplacées par respectivement $\widehat{\mu}_o = T_o$ et

$$\widehat{\sigma_o^2} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(x_i^{-1} - T_o\right)^2 . \tag{67}$$

Le bais estimé est donc donné par

$$\widehat{Biais}_2(\widehat{m}) = T_o^{-3}\, \widehat{\sigma_o^2}. \tag{68}$$

Il apparaît alors que la moyenne harmonique estimée $\widehat{m} = n\left(\sum_{i=1}^{n} x_i^{-1}\right)$ est un estimateur biaisé de la moyenne harmonique de la population: $\widehat{m}$ sous-estime m si la quantité T est négative et sur-estime m si T est positive. Un estimateur moins biaisé de m est par suite

$$\widehat{m}_c = T_o^{-1}\left(1 - T_o^{-2}\widehat{\sigma_o^2}\right). \tag{69}$$

4.4.2 Calcul de l'erreur type d'une moyenne estimée

L'erreur type $\sigma_{\widehat{m}}$ d'un estimateur $\widehat{m} = \psi^{-1}(T_o)$ est son écart type autour de sa valeur attendue, *i.e.*

$$\sigma_{\widehat{m}}^2 = E\{[\widehat{m} - E\{\widehat{m}\}]^2\} = E\{\widehat{m}^2\} - [E\{\widehat{m}\}]^2 . \tag{70}$$

Avec le développement limité de $\psi^{-1}(\cdot)$ à l'ordre 1 (52), $\widehat{m} - E\{\widehat{m}\} = \psi^{-1}(T_o) - \psi^{-1}(\mu_o) = g(\mu_o)(T_o - \mu_o)$. On a donc $\sigma_{\widehat{m}}^2 = [g(\mu_o)]^2\, E\{(T_o - \mu_o)^2\}$, ce qui donne

$$\sigma_{\widehat{m}}^2 = [g(\mu_o)]^2\, \sigma_o^2. \tag{71}$$

Étant donné que μ_o et σ_o^2 sont estimé sans biais par $\widehat{\mu}_o = T_o$ et $\widehat{\sigma_o^2} = n^{-1}\widehat{\sigma_T^2}$ avec

$$\widehat{\sigma_T^2} = \frac{n}{n-1} S_T^2. \tag{72}$$

où $S_T^2 = n^{-1}\sum_{i=1}^{n} [\psi(x_i) - T_o]^2$, la variance de $\widehat{m}$ est approximée à l'ordre 1 par

$$\sigma_{\widehat{m}}^2 = [g(T_o)]^2\, \widehat{\sigma_o^2}. \tag{73}$$

A partir du développement limité de $\psi^{-1}(\cdot)$ à l'ordre 2 (54), la variance approximative de $\widehat{m}$ est:

$$\sigma_{\widehat{m}_2}^2 \;=\; [g(\mu_o)]^2\sigma_o^2 + g(\mu_o)h(\mu_o)\delta_o^3 + [h(\mu_o)]^2\left[\frac{1}{4}\sigma_{T_o^2}^2 - \mu_o\left(\delta_o^3 + \mu_o\sigma_o^2\right)\right]. \tag{74}$$

où $\delta_o^3 = E\{(T_o - \mu_o)^3\}$ est le moment centré d'ordre 3 de T_o et $\sigma^2_{T_o^2} = E\{T_o^4\} - \left(E\{T_o^2\}\right)^2$ est la variance de T_o^2. Étant donnée que T_o est une moyenne arithmétique, ces moments sont estimés sans biais par

$$\widehat{\delta_o^3} = \frac{1}{n^2}\widehat{\delta_T^3} \quad \text{et} \tag{75}$$

$$\widehat{\sigma^2_{T_o^2}} = \frac{1}{n^3}\left[\widehat{\tau_T^4} + 4n\widehat{\mu_T\delta_T^3} + (2n-3)\widehat{\sigma_T^4} + 4n^2\widehat{\mu_T^2\sigma_T^2}\right] \quad \text{avec} \tag{76}$$

$$\widehat{\delta_T^3} = \frac{n^2}{(n-1)(n-2)}R_T^3 \tag{77}$$

$$\widehat{\sigma_T^4} = \frac{n^2(n^2-3n+3)}{n^3-5n^2+9n-3}\left[\frac{S_T^4}{n-1} - \frac{Q_T^4}{n^2-3n+3}\right] \tag{78}$$

$$\widehat{\tau_T^4} = \frac{1}{(n^2-3n+3)}\left[\frac{n^3}{n-1}Q_T^4 - 3(n-3)\widehat{\sigma_T^4}\right] \tag{79}$$

$$\widehat{\mu_T\delta_T^3} = \frac{1}{n^2-8n-30}\left[n(n-2)T_o\widehat{\delta_T^3} - 3n(n+5)T_o^2\widehat{\sigma_T^2} - \frac{n^2-5n-15}{n}\widehat{\tau_T^4} + \frac{3(2n^2-15)}{n}\widehat{\sigma_T^4}\right] \tag{80}$$

$$\widehat{\mu_T^2\sigma_T^2} = T_o^2\widehat{\sigma_T^2} - \frac{\widehat{\tau_T^4} + (n-3)\widehat{\sigma_T^4}}{n^2} - \frac{\widehat{\mu_T\delta_T^3}}{n} \tag{81}$$

où $R_T^3 = n^{-1}\sum_{i=1}^n [\psi(x_i) - T_o]^3$ et $Q_T^4 = n^{-1}\sum_{i=1}^n [\psi(x_i) - T_o]^4$. La variance de $\widehat{m}$ est ainsi approximé à l'ordre 2 par

$$\sigma^2_{\widehat{m}_2} = [g(T_o)]^2\widehat{\sigma_o^2} + g(T_o)h(T_o)\widehat{\delta_o^3} + [h(T_o)]^2\left[\frac{1}{4}\widehat{\sigma^2_{T_o^2}} - T_o\left(\widehat{\delta_o^3} + T_o\widehat{\sigma_o^2}\right)\right]. \tag{82}$$

qui se réduit à l'approximation à l'ordre 1 (73) si $h(T_o) = 0$ et à $\widehat{\sigma_o^2}$ si de plus $g(T_o) = 1$. Notons qu'après le calcul des quantités $\widehat{\sigma_T^4}$ (78) et $\widehat{\tau_T^4}$ (79), δ_o^3 est estimé sans biais par (75) et τ_o^4 est estimé sans biais par $\widehat{\tau_o^4} = \frac{1}{n^3}\left[\widehat{\tau_T^4} + 3(n-1)\widehat{\sigma_T^4}\right]$. Ainsi, les quantités requises pour la détermination d'une erreur type approximative à l'ordre 2 sont suffisante pour calculer le biais à l'ordre 3 (57) ou à l'ordre 4 (58).

7.4.2.1 Cas de la moyenne arithmétique

Dans le cas de l'estimation de la valeur attendue m du caractère X dans une population par la moyenne arithmétique $\widehat{m} = \bar{X}$ l'erreur type $\widehat{\sigma}$ de $\widehat{m}$ (écart type de la valeur estimée) satisfait $\sigma^2_{\widehat{m}} = \frac{\sigma^2}{n}$. On retrouve ce résultat en considérant une transformation linéaire $\psi(\cdot)$, *e.g.* $\psi(t) = t$, $g(t) = 1$ et $h(t) = 0$. L'erreur type de $\widehat{m}$ est alors estimée sans biais à travers $\sigma^2_{\widehat{m}} = \widehat{\sigma_o^2} = \frac{1}{n(n-1)}\sum_{i=1}^n (x_i - \bar{X})^2$.

7.4.2.2 Cas de la moyenne quadratique

Avec les dérivées $g(t) = \frac{1}{2}t^{-1/2}$ et $h(t) = -\frac{1}{4}t^{-3/2}$, la variance de $\widehat{m} = \sqrt{T_o}$ approximée à l'ordre 2 est

$$\sigma^2_{\widehat{m}} = \frac{\widehat{\sigma^2_o}}{4T_o} - \frac{3\widehat{\delta^3_o}}{16T_o^2} + \frac{1}{16T_o^3}\left[\frac{1}{4}\widehat{\sigma^2_{T_o^2}} - T_o^2\widehat{\sigma^2_o}\right]. \tag{83}$$

avec $T_o = n^{-1}\sum_{i=1}^n x_i^2$. Le premier terme de (83) correspond à la variance de $\widehat{m}$ approximée à l'ordre 1.

7.4.2.3 Cas de la moyenne géométrique

Ici, on $g(t) = h(t) = \exp(t)$. L'erreur type de $\widehat{m} = \exp(T_o)$ avec $T_o = n^{-1}\sum_{i=1}^n \log(x_i)$ satisfait donc

$$\sigma^2_{\widehat{m}} = \exp(2T_o)\widehat{\sigma^2_o} + \exp(2T_o)\left[\frac{1}{4}\widehat{\sigma^2_{T_o^2}} + (1 - T_o)\,\widehat{\delta^3_o} - T_o^2\widehat{\sigma^2_o}\right]. \tag{84}$$

Le premier terme de (84) correspond à la variance de $\widehat{m}$ approximée à l'ordre 1.

7.4.2.4 Cas de la moyenne harmonique

Les deux premières derivées de la transformation $\psi(\cdot)$ sont ici donnée par $g(t) = -t^{-2}$ et $h(t) = 2t^{-3}$. L'erreur type de $\widehat{m} = T_o^{-1}$ avec $T_o = n^{-1}\sum_{i=1}^n x_i^{-1}$ satisfait donc

$$\sigma^2_{\widehat{m}} = \frac{\widehat{\sigma^2_o}}{T_o^4} - \frac{6\widehat{\delta^3_o}}{T_o^5} + \frac{1}{T_o^6}\left[\widehat{\sigma^2_{T_o^2}} - 4T_o^2\widehat{\sigma^2_o}\right], \tag{85}$$

le premier terme de (85) correspondant à la variance de $\widehat{m}$ approximée à l'ordre 1.

4.4.3 Calcul de l'erreur quadratique attendue d'une moyenne estimée

L'erreur quadratique attendue (EQA) d'un estimateur $\widehat{m}$ est sa variance autour de la vraie valeur m, *i.e.* qu'elle satisfait

$$\varepsilon^2_{\widehat{m}} = E\{(\widehat{m} - m)^2\}. \tag{86}$$

L'EQA exprime l'incertitude totale sur une estimation et est se décompose comme suit:

$$\begin{aligned}\varepsilon^2_{\widehat{m}} &= [E\{\widehat{m}\} - m]^2 + E\{\widehat{m}^2\} - [E\{\widehat{m}\}]^2\,.\\ \varepsilon^2_{\widehat{m}} &= [Biais(\widehat{m})]^2 + \sigma^2_{\widehat{m}}.\end{aligned} \tag{87}$$

Dans un ensemble d'estimateurs d'une même quantité, le meilleur estimateur est celui d'EQA minimum. Ainsi, si deux estimateurs ont le même biais, celui ayant la plus faible variance (dit le plus efficient) est le meilleur. De même, un estimateur biaisé peut être préféré à un estimateur non biaisé si la variance de l'estimateur biaisé est si petite que son incertitude global (EQA) est inférieure à la variance de l'estimateur non biaisé. L'estimateur non biaisé de variance minimum

est généralement considéré comme le meilleur estimateur d'une quantité. Mais il n'est pas facile à trouver!

4.4.4 Calcul des limites de confiance autour d'une moyenne estimée

Le calcul des limites de confiance autour d'une moyenne estimée suppose la connaissance de la distribution de l'estimateur considéré. Néanmoins, le théorème centrale limite permet de construire un intervalle de confiance asymptotique autour d'une moyenne arithmétique d'une série, quelle que soit la distribution de la variable d'intérêt considérée. Ainsi, puisque les moyennes considérées ici sont des transformations de moyennes arithmétiques, on peut obtenir leurs limites de confiance en transformant les limites de confiance autour des moyennes arithmétiques de base. Cette approche permet de garder les limites de confiance dans le domaine de la variable de départ, contrairement à une approche basée sur l'application du théorème centrale limite développé pour les moyennes régulières (de Carvalho, 2016).

7.4.4.1 Cas de la moyenne arithmétique

Ici, la moyenne estimée est $\widehat{m} = \bar{X}$. Le rapport $\frac{\widehat{m}-\mu}{\sigma/\sqrt{n}}$ suit une distribution normale standard. Si la variance σ^2 de X est connue, un intervalle de confiance au niveau $100(1-\alpha)\%$ est donné par

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}}\right] \tag{88}$$

où $Z_{1-\frac{\alpha}{2}}$ est le percentile d'ordre $1-\frac{\alpha}{2}$ de la distribution normal standard. Si $\alpha = 5\%$, $Z_{97.5\%} \approx 1.95996$ généralement arrondi à $Z_{97.5\%} \approx 1.96 \approx 2$. Mais en pratique, σ^2 est souvent inconnue et la variance $\frac{\sigma^2}{n}$ de $\bar{X}$ est estimée par $\widehat{\sigma_o^2} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{X})^2$. Dans ce cas, le rapport $\frac{\widehat{m}-\mu}{\widehat{\sigma_o}}$ avec $\widehat{\sigma_o} = \sqrt{\widehat{\sigma_o^2}}$ suit une distribution t de Student de $n-1$ degrés de liberté. Un intervalle de confiance au niveau $100(1-\alpha)\%$ est alors donné par

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - \widehat{\sigma_o} t^{n-1}_{1-\frac{\alpha}{2}}, \bar{X} + \widehat{\sigma_o} t^{n-1}_{1-\frac{\alpha}{2}}\right] \tag{89}$$

où $t^{n-1}_{1-\frac{\alpha}{2}}$ est le percentile d'ordre $1-\frac{\alpha}{2}$ de la distribution t de Student de $n-1$ degré de liberté. Si *e.g.* $\alpha = 5\%$, on a respectivement pour des tailles $n = 5, 10, 30, 50, 75, 100, 500, 1000$, $t^{4}_{97.5\%} \approx 2.77645$, $t^{9}_{97.5\%} \approx 2.2622$, $t^{29}_{97.5\%} \approx 2.04523$, $t^{49}_{97.5\%} \approx 2.00958$, $t^{74}_{97.5\%} \approx 1.99254$, $t^{99}_{97.5\%} \approx 1.98422$, $t^{499}_{97.5\%} \approx 1.96473$ et $t^{999}_{97.5\%} \approx 1.96234$.

7.4.4.2 Cas des moyennes quasi-arithmétiques

Dans les cas de moyennes quasi-arithmétiques (*e.g.* quadratique, géométrique, harmonique) ou $\widehat{m} = \psi^{-1}(T_o)$, l'intervalle de confiance au niveau $100(1-\alpha)\%$ autour de μ_T est

$$IC_{1-\alpha}(\mu_T) = \left[T_o - \widehat{\sigma_o} t_{1-\frac{\alpha}{2}}^{n-1}, T_o + \widehat{\sigma_o} t_{1-\frac{\alpha}{2}}^{n-1}\right] \tag{90}$$

Un intervalle de confiance au niveau $100(1-\alpha)\%$ autour de la moyenne m est alors

$$IC_{1-\alpha}(m) = \left[\psi^{-1}\left(T_o - \widehat{\sigma_o} t_{1-\frac{\alpha}{2}}^{n-1}\right),\ \psi^{-1}\left(T_o + \widehat{\sigma_o} t_{1-\frac{\alpha}{2}}^{n-1}\right)\right] \quad \text{si } \psi^{-1}(\cdot) \text{ est croissante et} \tag{91}$$

$$IC_{1-\alpha}(m) = \left[\psi^{-1}\left(T_o + \widehat{\sigma_o} t_{1-\frac{\alpha}{2}}^{n-1}\right),\ \psi^{-1}\left(T_o - \widehat{\sigma_o} t_{1-\frac{\alpha}{2}}^{n-1}\right)\right] \quad \text{si } \psi^{-1}(\cdot) \text{ est decroissante.} \tag{92}$$